Sequencing Technology and Study Design

Michael C. Zody, Ph.D.

Workshop on Genomics

Cesky Krumlov

January 7, 2020

Logistics

- Introduction
- Please feel free to ask questions at any point
- Slides will be posted on workshop website
- One break at about 60 minutes

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing
- General Study Design Considerations
- Considerations for Specific Sequencing Assays

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing
- General Study Design Considerations
- Considerations for Specific Sequencing Assays

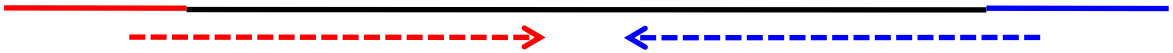
What is a read? What is a library?

- Definition of “read”: A single sequence from one fragment in the sequencing library (one cluster, bead, *etc*.)
- If generating paired reads, then 2 reads derived from each fragment in the library
- Definition of “library”: A collection of DNA fragments that have been prepared to be sequenced
- Definition of “coverage”: The number of reads spanning a particular base in the genome

Types of reads

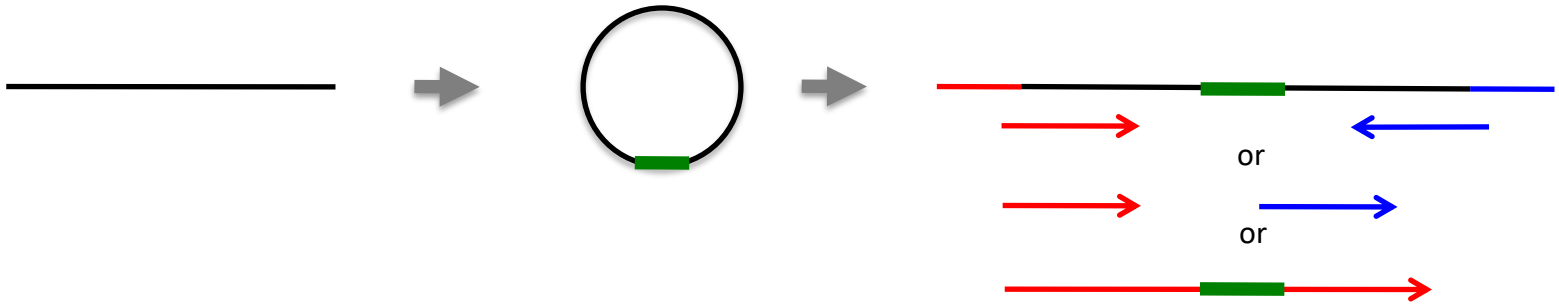
- Fragment reads (come from fragment libraries)
  - Single read in one direction from a fragment

- Paired end reads (come from fragment libraries)
  - Two reads from opposite ends of the same fragment
  - Reads point towards each other



Types of reads

- Mate Pair Reads (come from Jumping Libraries)
  - Long fragment of DNA is circularized
  - Junction is captured (e.g., by biotinylated adapter)
  - Remainder is cleaved (many methods)
  - Ends are sequenced
  - Read orientations depend on the exact method



Types of reads

- Linked reads (10X Genomics and others)
  - Long (50-100kb) DNA isolated in emulsion
  - Read pairs generated within the emulsion
  - Reads have an emulsion-specific barcode
  - Sequence normal read pairs
  - Can use reads normally for alignment/assembly
  - Can also group reads by haplotype of origin

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing
- General Study Design Considerations
- Considerations for Specific Sequencing Assays

Sanger Sequencing (1977)

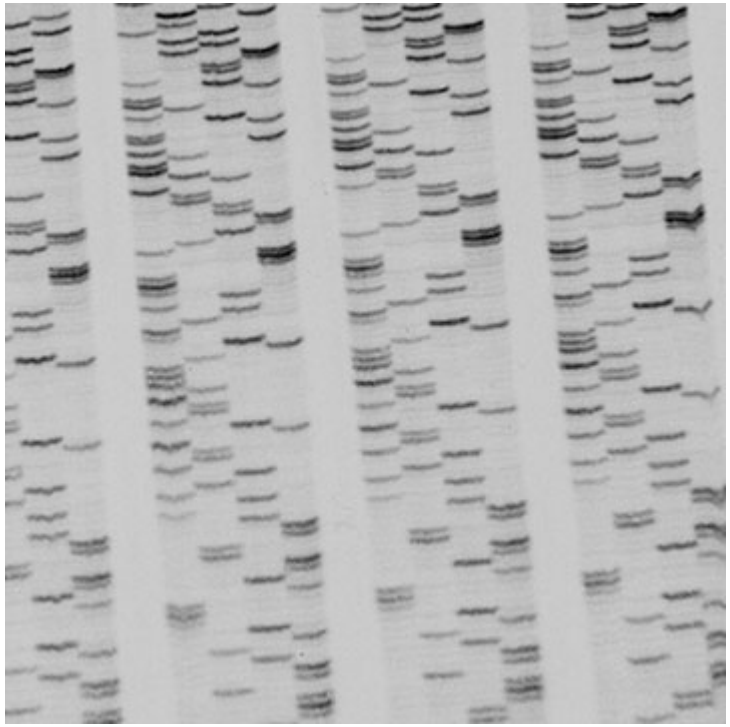
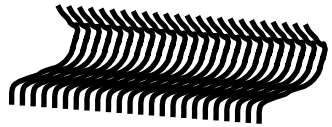
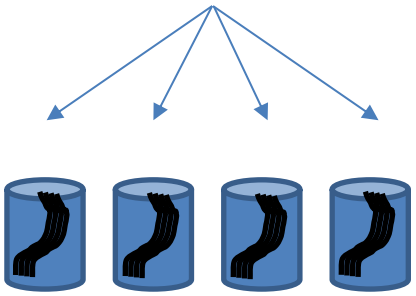


Image credit: https://unlockinglifescode.org/timeline/11

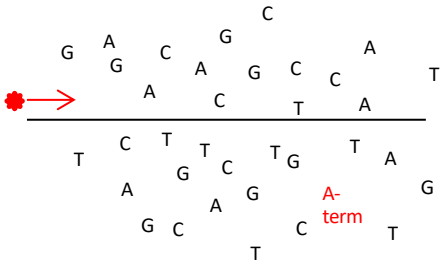
How Sanger Sequencing Works



Single-stranded DNA



Split into four reactions



Add labeled primer and nucleotides, one of which is sometimes terminated

Allow the strands to extend

A	T	A	T	A	T
G	C	G	C	G	C
C	G	C	G	C	G
T	A	T	A	T	A
G	C	G	+	G	C
A	T	A		A	T
C	G	C		C	G
T	A	T		T	A
A	T	A		A	+
G	C	G		G	
A	T	A		A	
T	A	T		T	
C	+	C		C	
G		G		G	

Automation of Sanger (1986)

- Replacement of radioactive label with laser-excitable fluorescent dyes
- Allowed all four nucleotides to be run in a single lane of a gel
- Base sequence could be read off with a camera as the fluorescing strands passed a certain point near the end of the gel
- Signal from each lane could be converted to a nucleotide sequence by a computational process called base calling

Fluorescent slab gel image

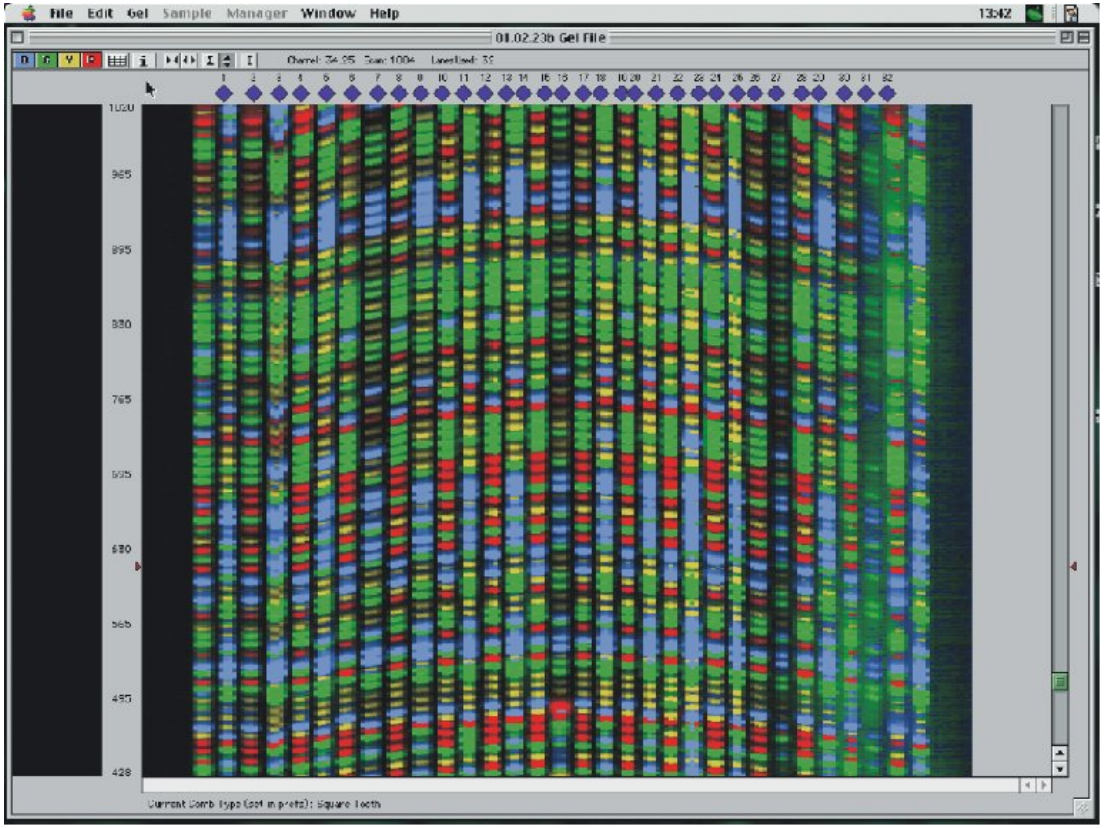


Image credit: http://www.mun.ca/biology/scarr/How_it_works.htm

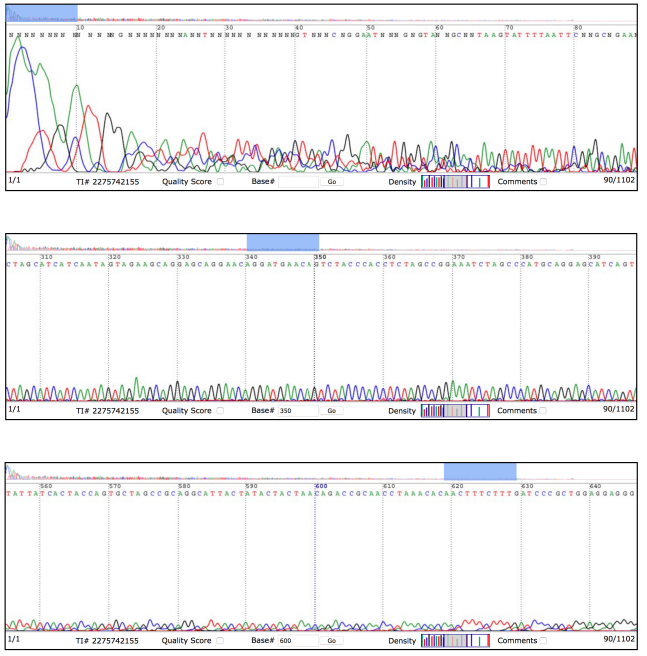
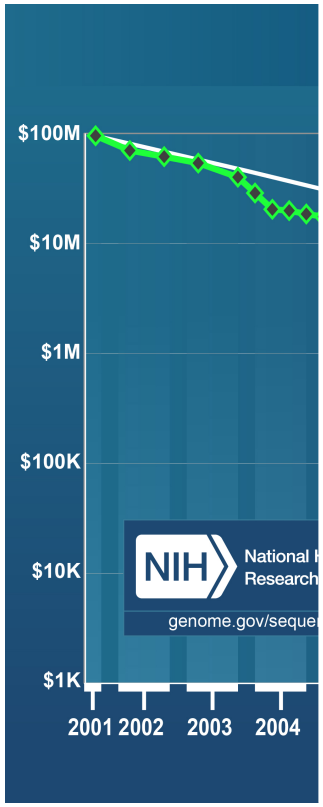


Image credit: NCBI Trace Archive

Capillary Gel Sequencing (1998)

- Replacement of 2D slab gels with an array of enclosed capillaries
- Cleaner signal processing
- Fully automated loading
- Faster run times

Cost Curve for Sanger Gel Sequencing



Other Early Technologies

- Maxam-Gilbert sequencing
- LI-COR
- Molecular Dynamics MegaBACE
- Pyrosequencing
- Mass spectrometry

Statistics of Apex Capillary Sequencing

- 96 reads per run
- 700-1000 bases per read
- Very high base accuracy over most of the read length (<1/100,000)
- ~$1 per read
- ~1 run per hour
- ~2 million bases per machine per day
- Large sequencing centers could do a single mammalian genome to assembly depth in about 2-3 months

Limits on Sanger Gel Performance

- Tradeoff between loss of signal due to diffusion and loss of resolution at high voltage or short gel length
- Longer gels/capillaries or slower voltages provide better separation of short to medium fragments
- Longer gel run times mean more diffusion of fragments in the gel, which blurs adjacent signal and spreads peaks
- The maximal high quality read length is around 1000 bp

454 Sequencing

- First “Next Generation” or massively parallel technique
- Based on pyrosequencing
- Emulsion PCR DNA prep on beads
- Beads loaded into a picotiter plate for sequencing

Emulsion PCR

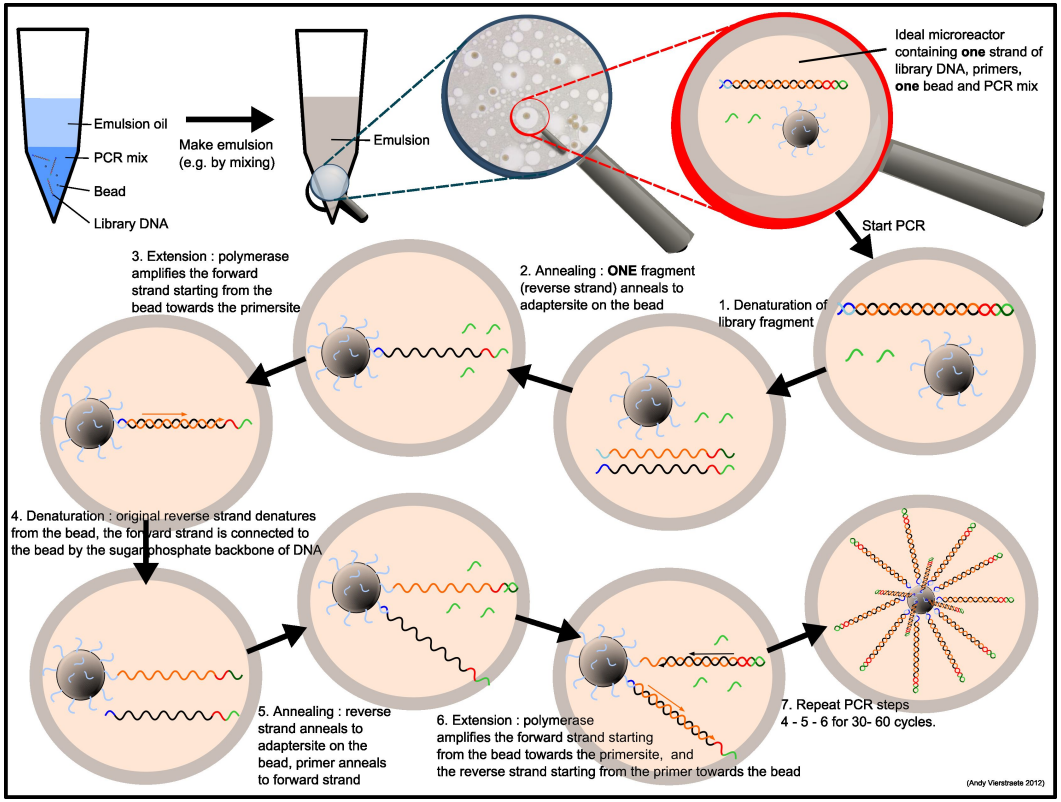
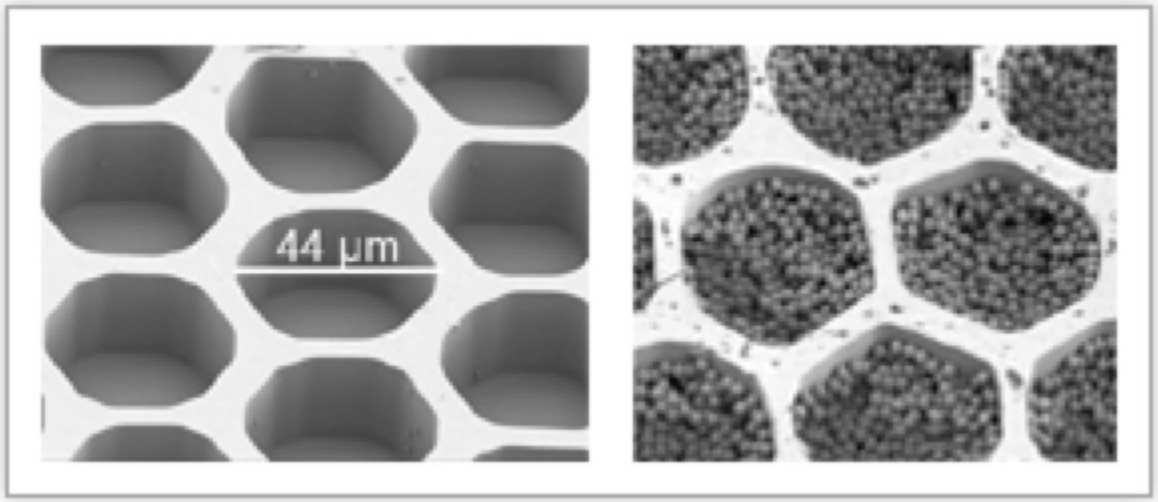


Image credit:
https://users.ugent.be/~avierstr/nextgen/nextgen.html

454 Picotiter Plate



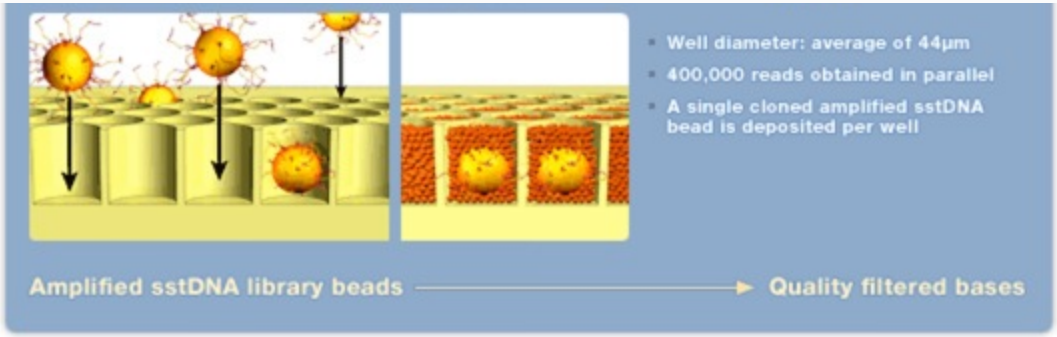
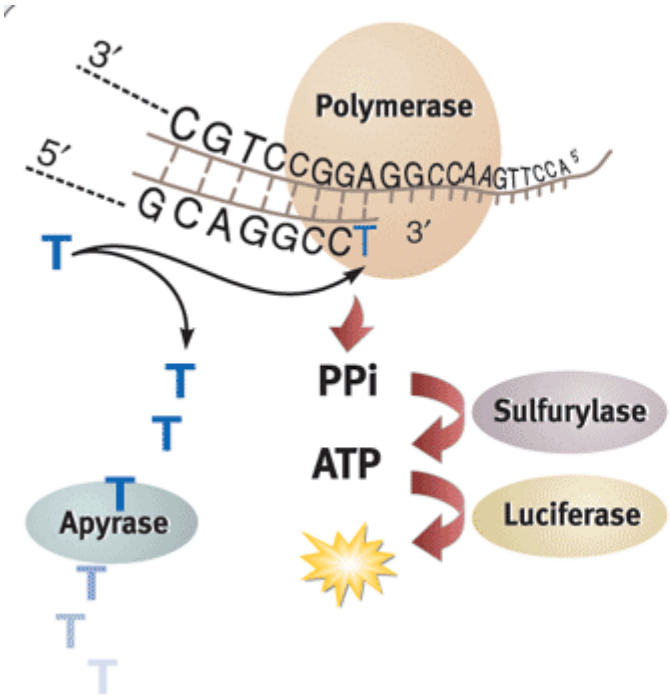
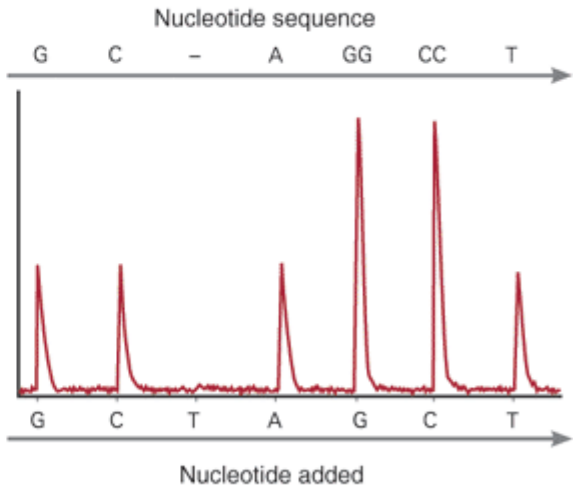


Image credit:
http://www.mbio.ncsu.edu/MB451/lectureModules/molecularEcology/molecularSurveys/454/454.html

Pyrosequencing





454 Output: the Flowgram

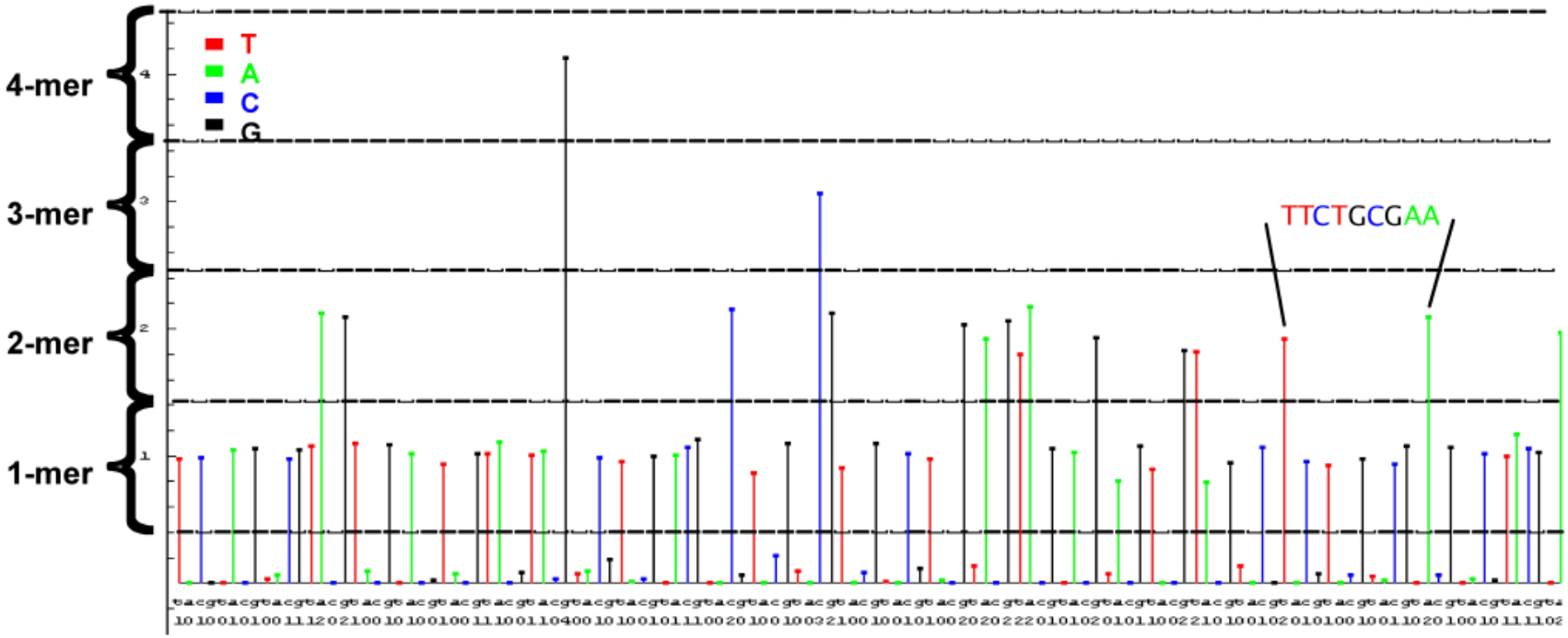
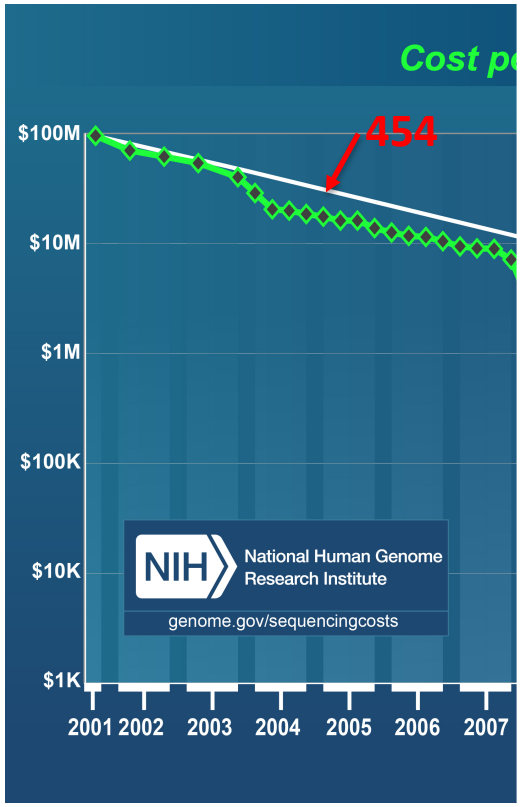


Image credit: https://contig.wordpress.com/2010/10/28/newbler-input-i-the-sff-file/

Cost Curve for 454 Sequencing



Statistics of Apex 454 Sequencing

- 1 million reads per run
- 400-500 bases per read (750?)
- High error rate (~1.5%), very motif dependent (homopolymers)
- Cost several thousand dollars per run
- ~10 hours per run
- ~1 billion bases per machine per day

Limits on 454 Performance

- Failure to accurately read homopolymers or sequences near homopolymers; physical limit on ability to read the full incorporation
- Loss of signal over time
- Signal/noise degradation due to asynchrony of extending strands
- Length of fragment that could be amplified on bead in emPCR
- No ability to sequence the second strand of DNA or do non-contiguous reads

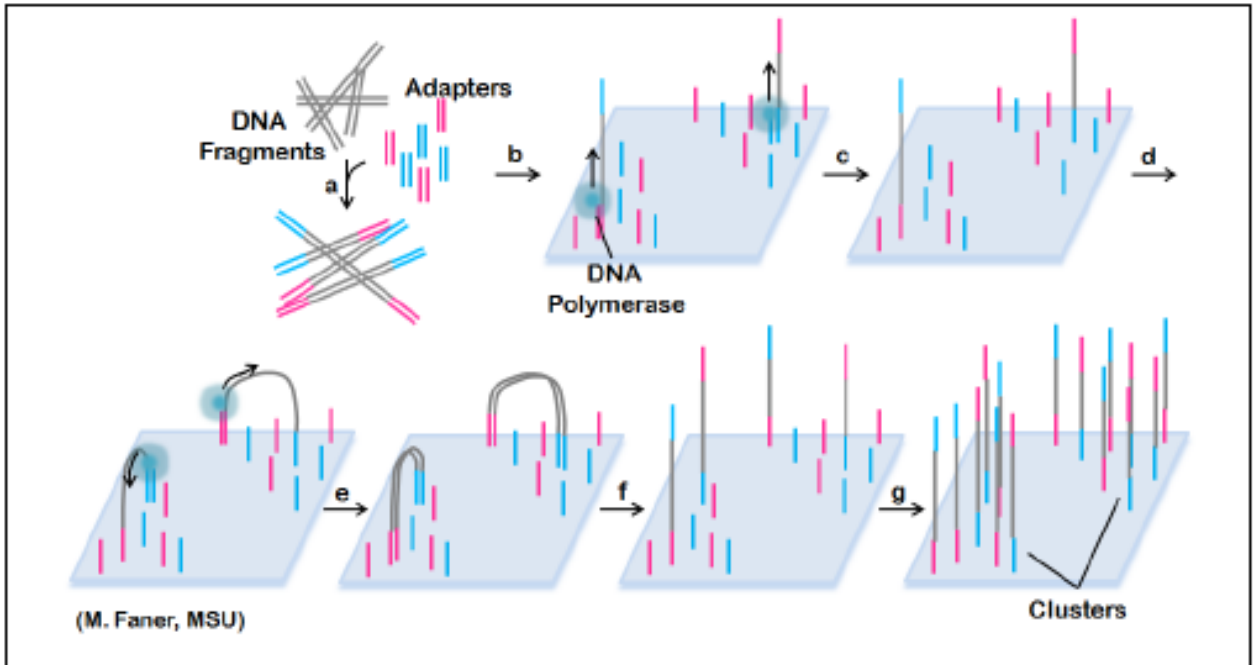
Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing
- General Study Design Considerations
- Considerations for Specific Sequencing Assays

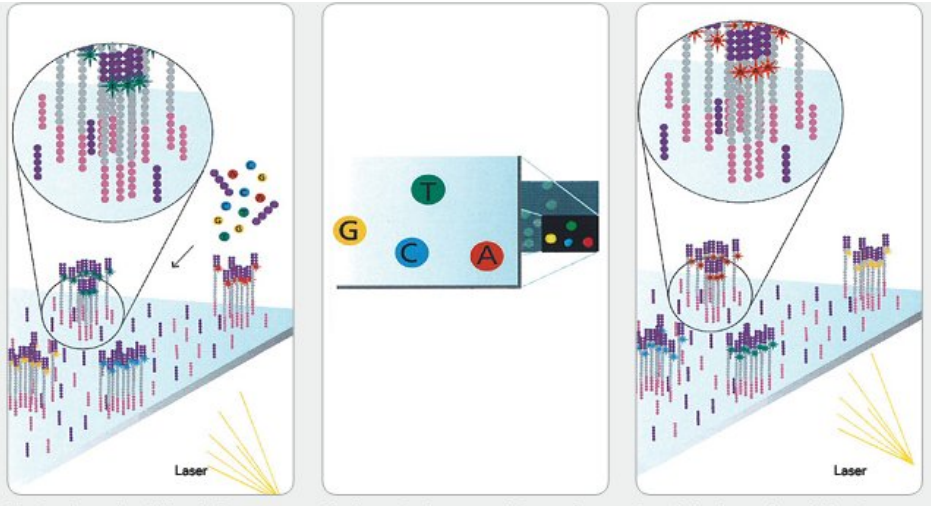
Illumina (Solexa) Sequencing

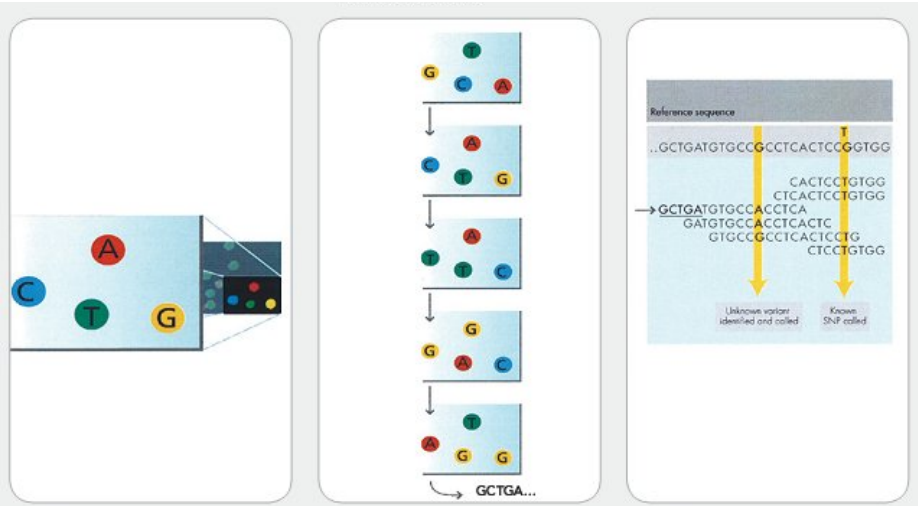
- Sequencing of DNA strands amplified *in situ* on a glass slide
- Use reversible terminators to sequence one base at a time

Bridge Amplification



Reversible Terminator Sequencing

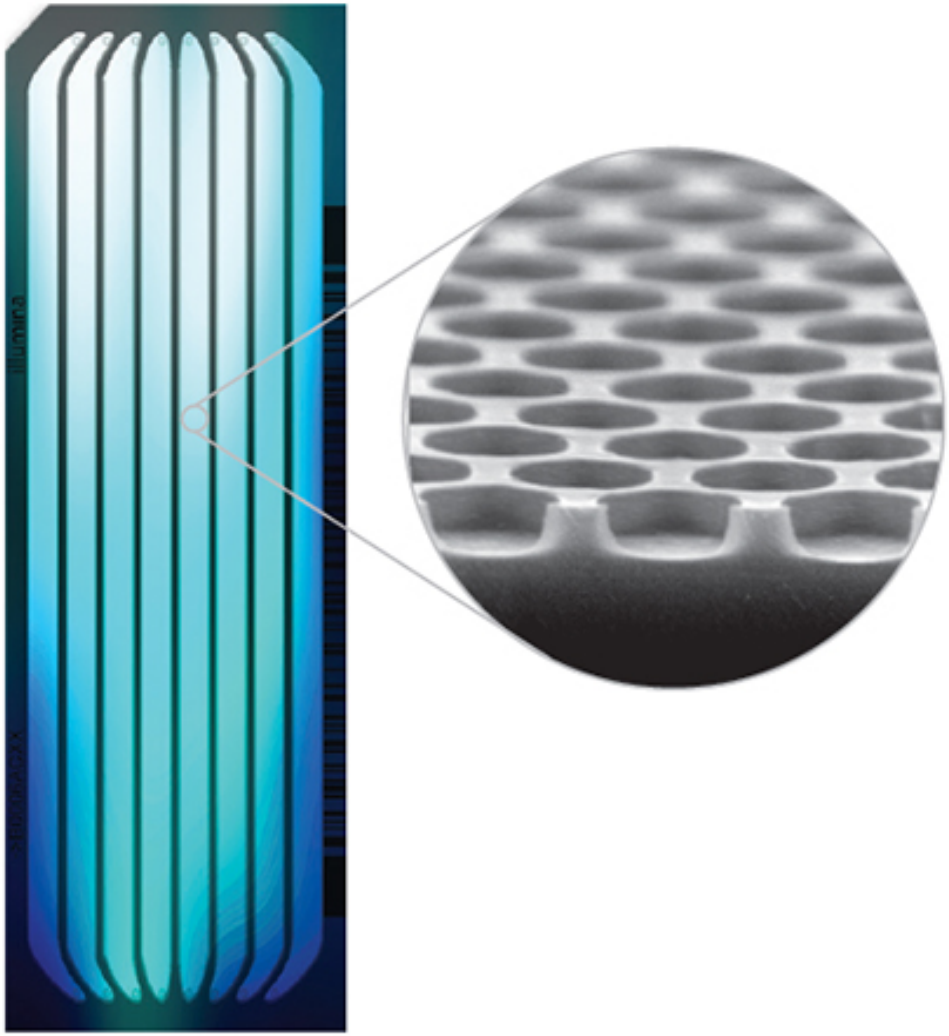




Recent Changes in Illumina

- Patterned flowcells
- Exclusion amplification
- 2 color chemistry

Patterned Flowcells

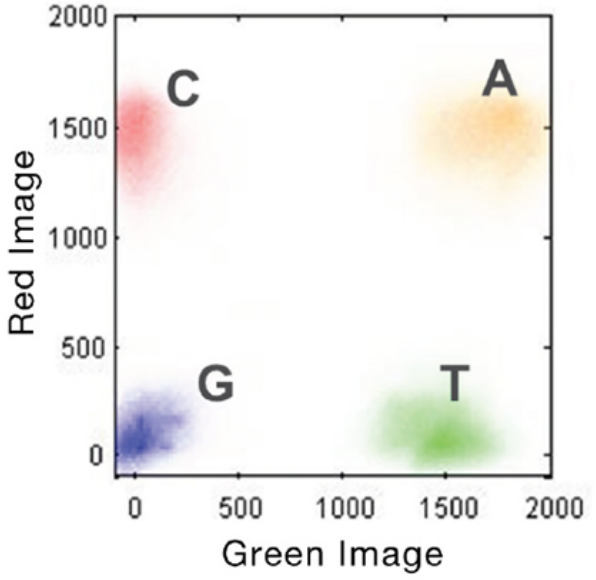


Exclusion amplification

- No more bridge PCR on patterned flowcells
- Fragments rapidly amplify as soon as they arrive at the patterned spot
  - Prevents a second fragment from amplifying there
  - Allows overloading to maximally fill flowcell
  - (Not perfect)
- Exact method is not described (see patent)
- Results in “proximal duplicates” or “pad hops”
- Problems with “index switching”

2 color chemistry

- One base (A) labeled with 2 colors
- One base (G) unlabeled
- Allows faster image scanning
- Dead clusters look like runs of G
  - Mostly do not align (in human)
  - ”Supplemental” alignments



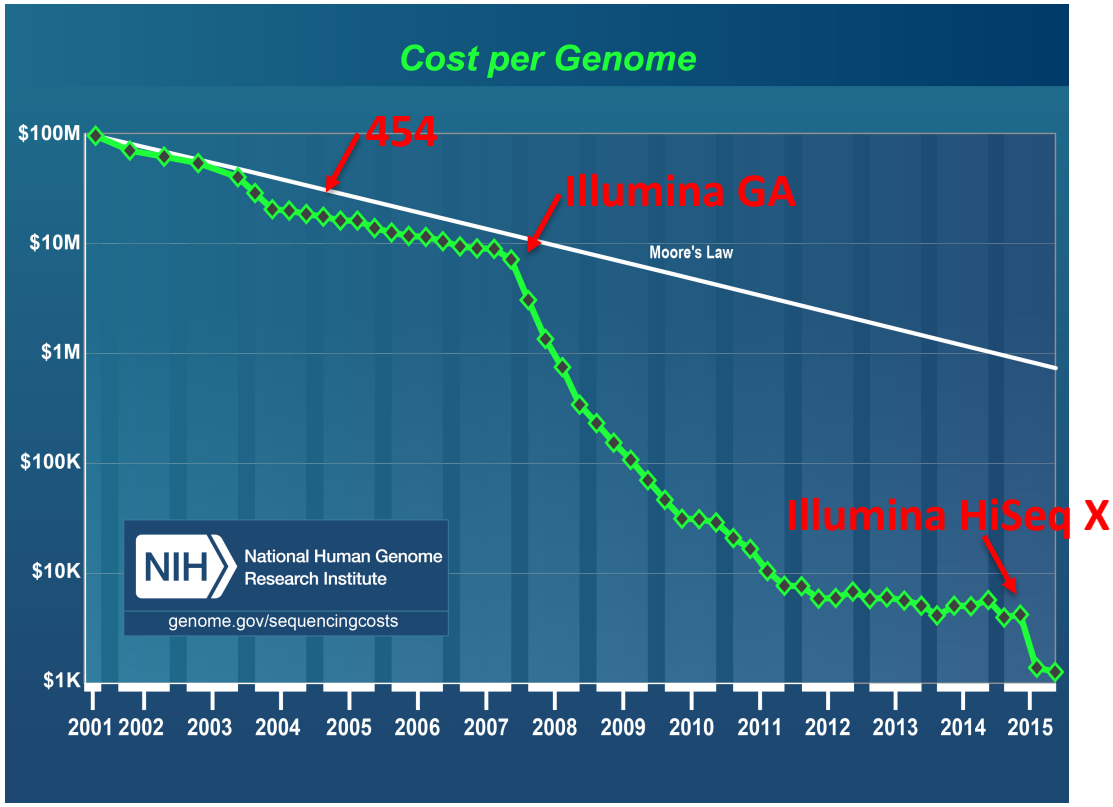
Statistics of Apex (so far) Illumina

- 10 billion read pairs per run
- 300 bases per read pair
- Relatively low error (<1%), some context dependency
- Cost ~$30,000 per run (for NovaSeq on largest flow cell size)
- ~2-3 days per run
- ~1 trillion bases per machine per day

Limits on Illumina Performance

- Loss of signal over time
- Signal/noise degradation due to asynchrony of extending strands
- Viability of sequencing reagents over the course of a run
- Length of fragment that could be amplified into clusters on the slide

Cost Curve for Illumina Sequencing



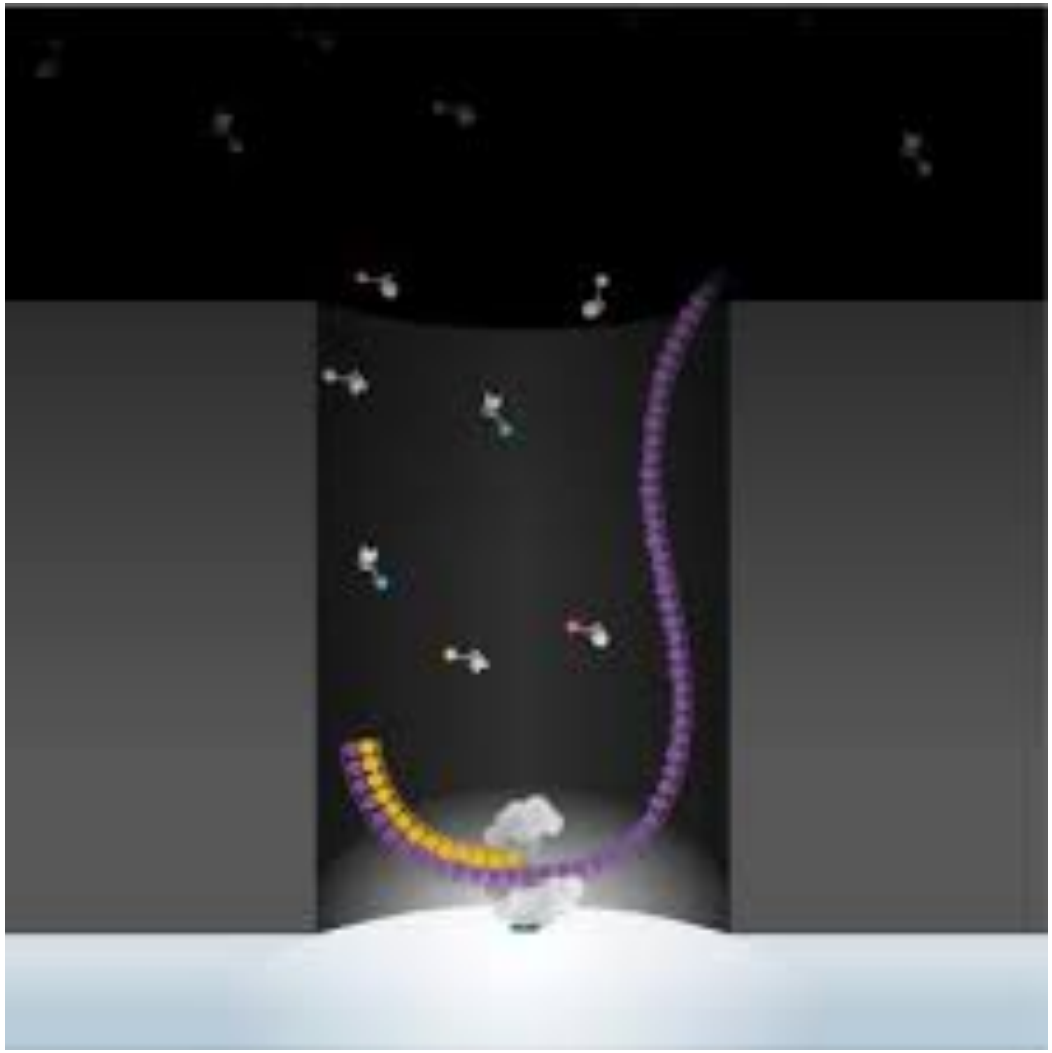
Other Next Generation Technologies

- SOLiD
  - Ligation rather than polymerase based
  - Used redundant base sampling with error correction ("color space") to enhance error rate (<0.1%), but made analysis very challenging
  - Short reads, limited second read capability
- Ion Torrent
  - Like 454 (emPCR, well-based sequencing)
  - Uses direct measurement of pH changes with base addtion ("post-light")
- Helicos
  - Like Illumina but with single molecules

Single Molecule (Third Generation) Methods

- Pacific Biosciences SMRT (Single Molecule Real Time) sequencing
  - Uses a polymerase anchored in a zero-mode waveguide
  - Images all wells at the same time in real time with digital video
  - Interprets bases by the light signal visible at incorporation
  - Very large instrument
- Oxford Nanopore Technologies
  - Uses protein nanopores in synthetic membrane to thread DNA through
  - Current sensors measure change in fluid current flow through pore to differentiate groups of multiple bases as they occupy the pore
  - Very small instrument (attaches to compute like a USB drive)

PacBio SMRT



https://www.youtube.com/watch?v=WMZmG00uhwU&feature=youtu.be

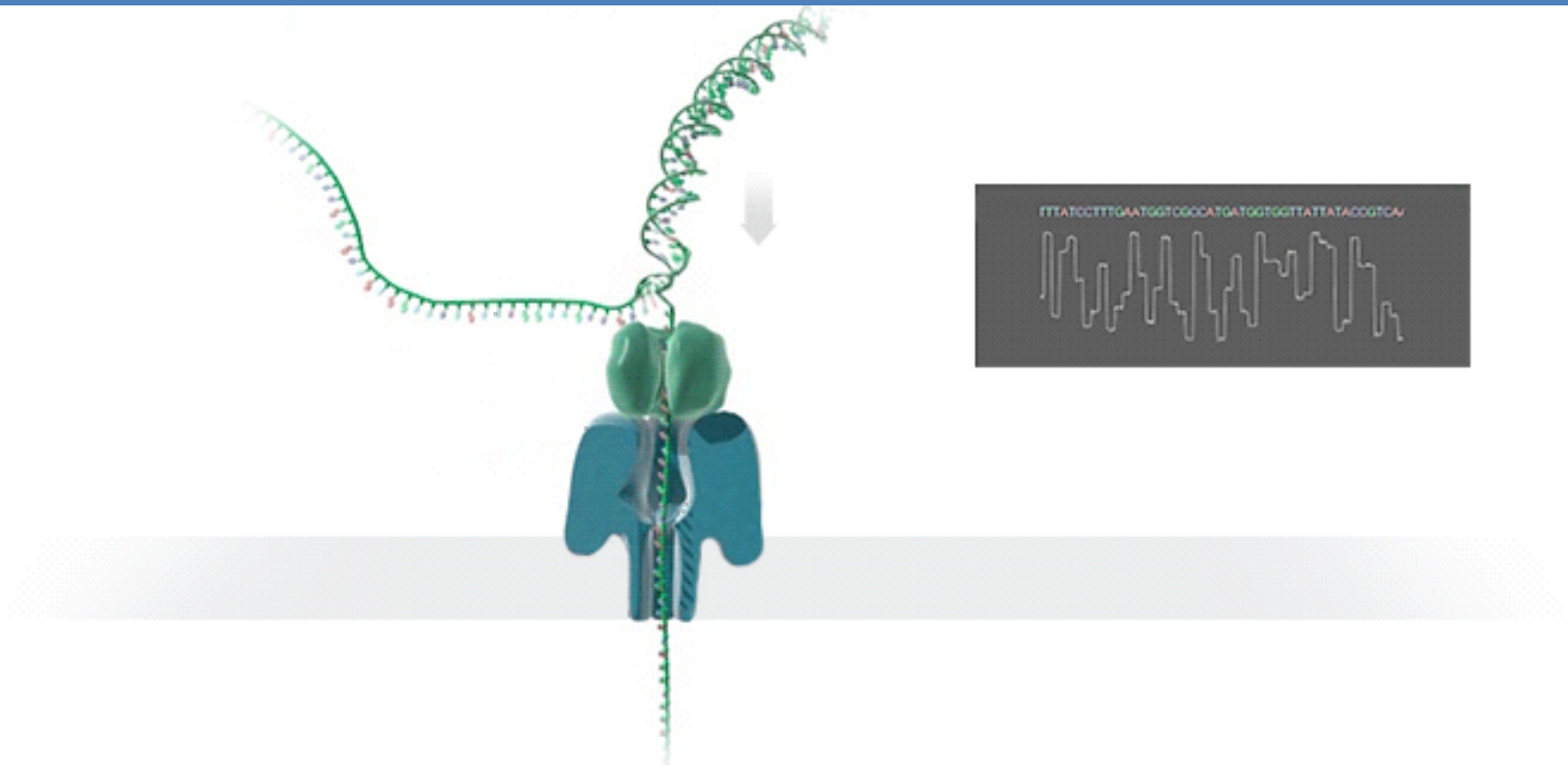
Current State of PacBio Sequencing

- Very long reads (50,000+)
- Very high error rates (15%+), but random
- Circular consensus sequencing high accuracy
- Low throughput per run
- Relatively short run times
- Much higher costs per base than Illumina (3-10-fold)
- Reads end when polymerase dies

PacBio “Hi-Fi” Reads

- Sized libraries, 10-20kb long
- Generate circular consensus on these
- Can read each linear piece 7-10 times
- Generates very high accuracy long reads
- Can assemble easily and even distinguish highly similar repeats
- Cost to generate deep coverage is still 2-3x higher than PacBio continuous long reads

Oxford Nanopore



Current State of Oxford Nanopore

- Very long reads (100,000+, >1 Mbp?)
- Very high error rates (15%+), non random
- Can read both strands to improve accuracy
- Low throughput
- Relatively short run times
- Much higher costs per base than Illumina (maybe 3-20-fold)
- Reads end when pores die

Limitations of Single Molecule Techniques

- Single molecule means no redundancy, so error rates will be high unless the same molecule can be read more than once
- Methods of detection are hard to massively parallelize
- Currently, these techniques actually require large amounts of DNA
- Getting very long reads requires very high quality input DNA
- Reads with higher error rates are more computationally expensive to align and assemble

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing
- General Study Design Considerations
- Considerations for Specific Sequencing Assays

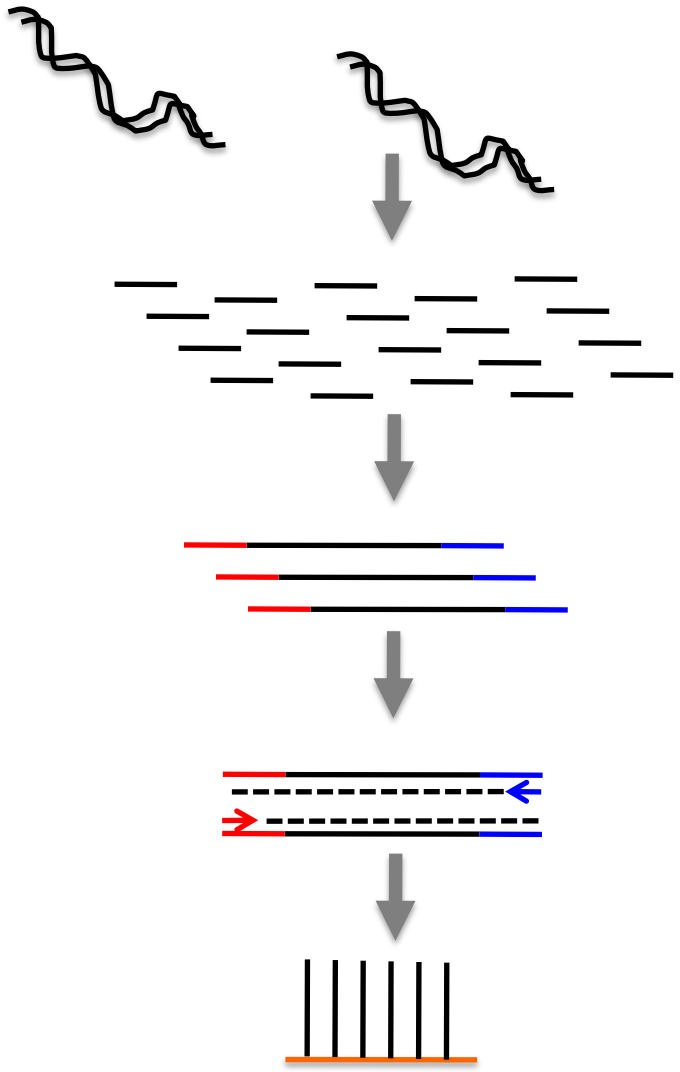
Prepping DNA for Sequencing

- Steps of library construction and sequencing
- Making Fragment libraries (to generate fragment or paired end reads)
- Making Jumping libraries (to generate mate pair reads)
- Pooling with or without barcoding
- Possible artifacts of library construction
  - PCR-based artifacts
  - Sequencing of primers, adapters, and tags

Steps of Library Construction

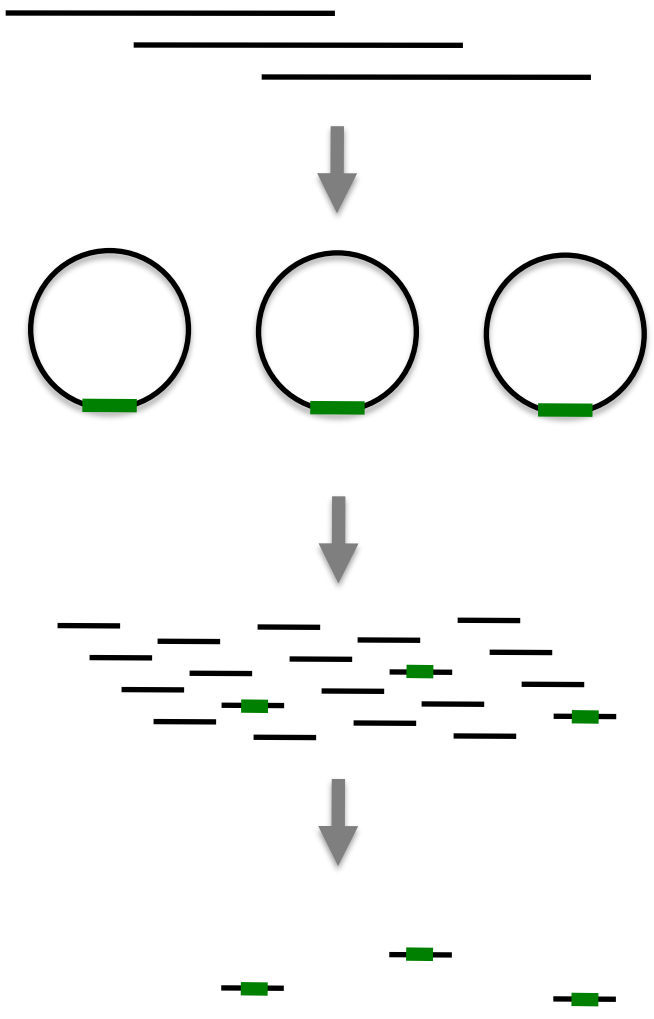
- Add adapters containing:
  - Barcodes (for multiplexing)
  - Sequencing primers
  - Amplification primers
  - Sequence for substrate attachment
- Amplify fragments by universal PCR
- Optionally pool barcoded libraries

Steps of Fragment Library Construction



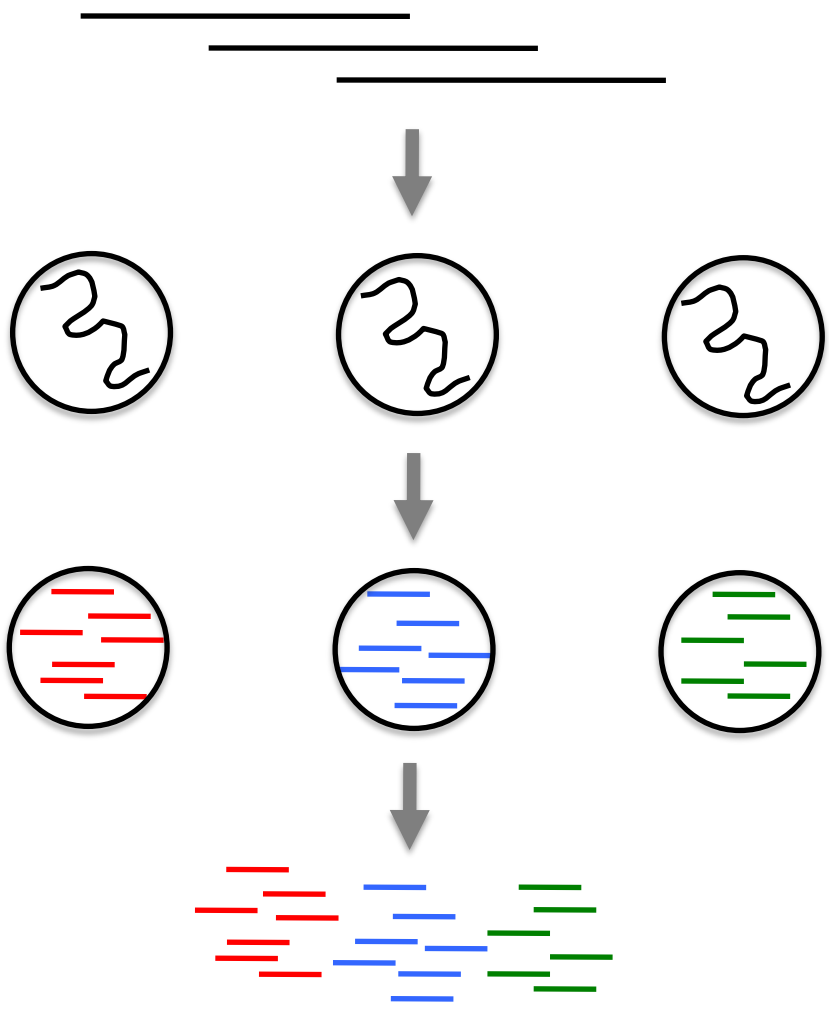
- Extract DNA
- Fragment and possibly size select (300-600 bp)
- Add adapters
- Amplify
- Select single molecules
- Amplify in clusters/beads

Steps of Jumping Library Construction



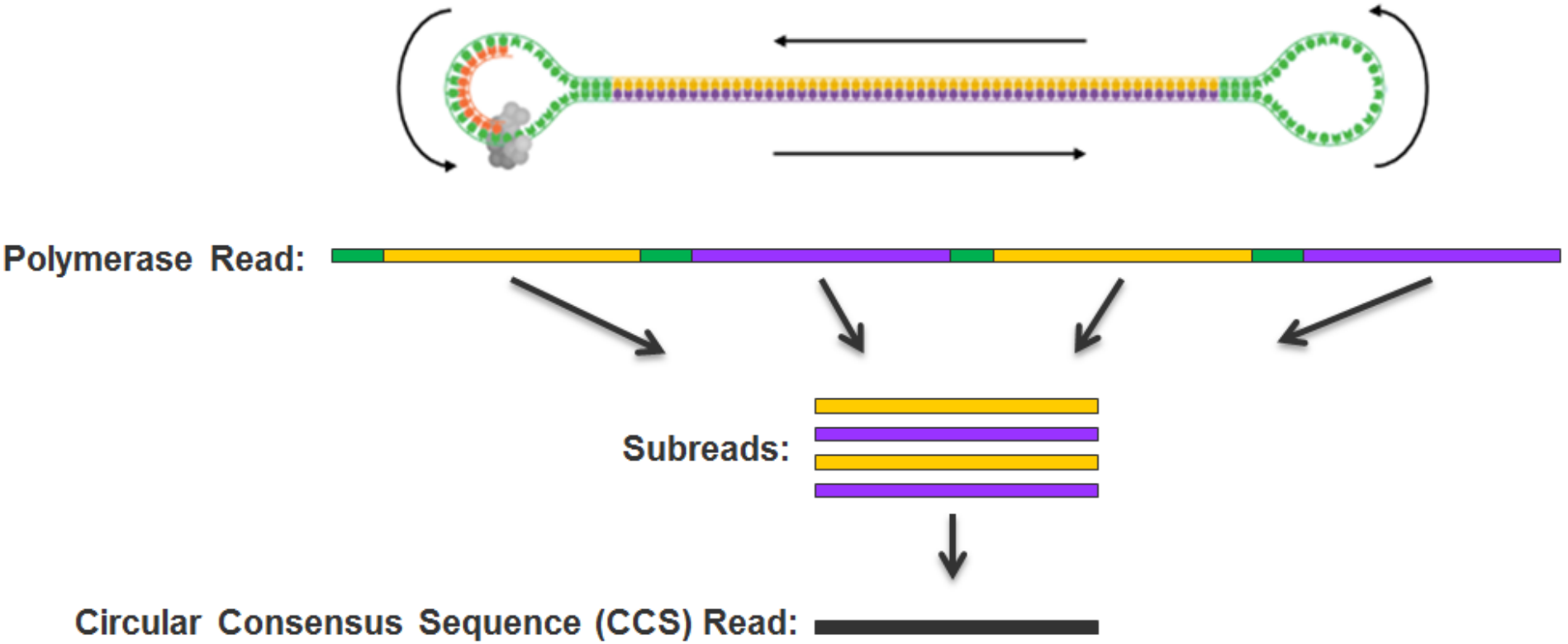
- Extract DNA, fragment and size select (2-40 kb)
- Circularize with labeled adapters
- Fragment and size select (300-600 bp)
- Select fragments containing labeled adapters
- Proceed as for fragment library

Steps of Linked Read Library Construction



- Extract DNA, fragment and size select (50+ kb)
- Isolate large fragments
- Fragment, barcode, and size select (300-600 bp)
- Pool and proceed as for fragment library

Steps of PacBio Library Construction



- Extract DNA, fragment and size select (50+ kb)
- Add hairpin adapters to both ends

Pooling with barcoding

- Unique DNA tags identify samples
- Allows multiple distinct samples on one run/lane
- Advantages:
  - Reduced cost of sequencing for small samples
  - Analysis is identical to unpooled data
- Disadvantages:
  - Some small throughput loss due to barcode fails
  - Data mis-assignment from bad barcode reads
  - Increased per sample cost for library construction

Pooling without barcoding

- Mix input DNA without identification
- No way to definitively separate data from different samples afterwards
- Advantages:
  - Single library prep for a number of samples
  - No yield lost to barcodes
- Disadvantages:
  - Loss of all individual associations
    - Loss of ability to use replicates!
  - No check on accuracy of pooling

PCR-based artifacts

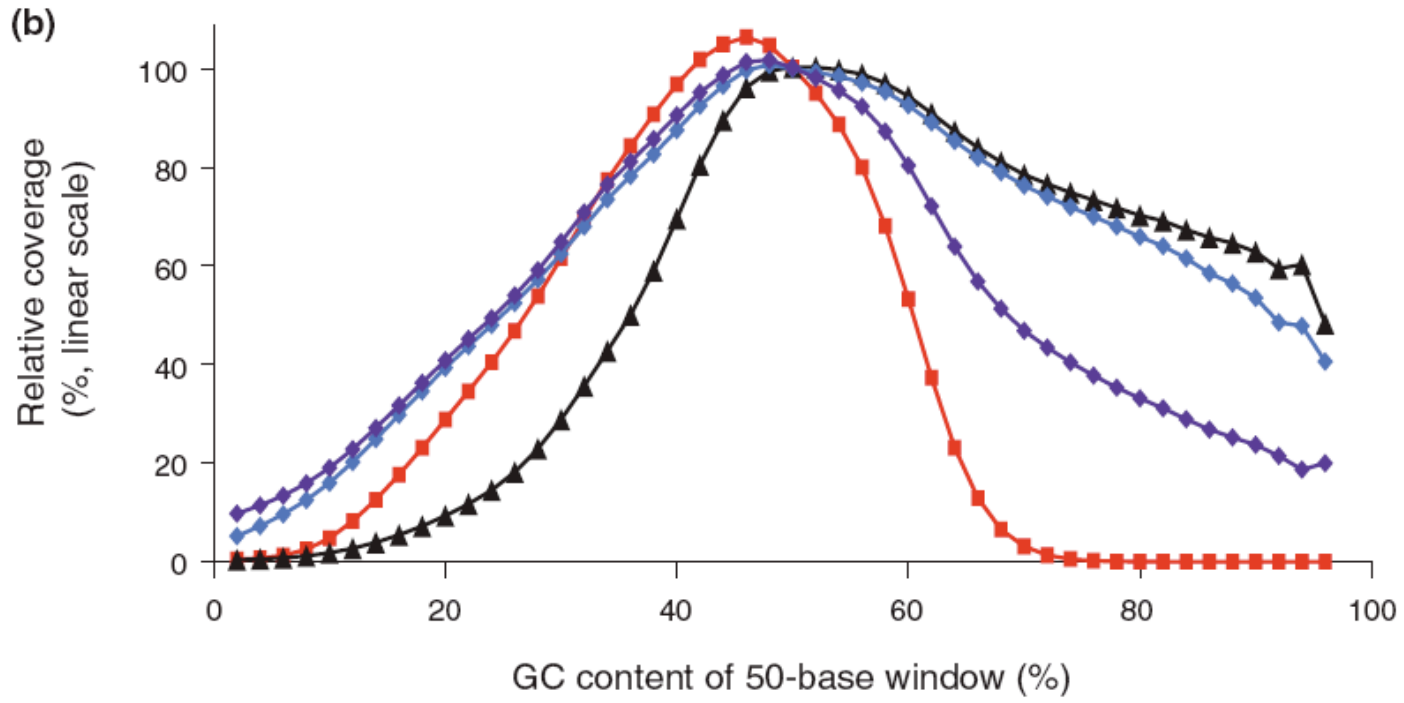
- Most libraries are PCR amplified during construction
- After library construction, single molecules are isolated and then amplified again for sequencing
- Errors from library construction PCR will not be detectable as sequencing errors
- Regions with secondary structure or extreme GC content:
  - Will amplify poorly and be underrepresented
  - May form small or weak clusters with poor sequence quality
- PCR may form chimeric sequences (especially in targeted designs)
- PCR amplification may result in duplicated sequences

PCR Errors: How Much PCR?

- You may be doing more PCR than you think
- Initial amplification of sample
- Targeting PCR
- Library amplification
- 100 rounds of PCR is equivalent to a 2 order of magnitude drop in polymerase accuracy

PCR-based artifacts: PCR bias

- Most PCR protocols work best for ~50% GC
- Extreme GC sequences are underrepresented



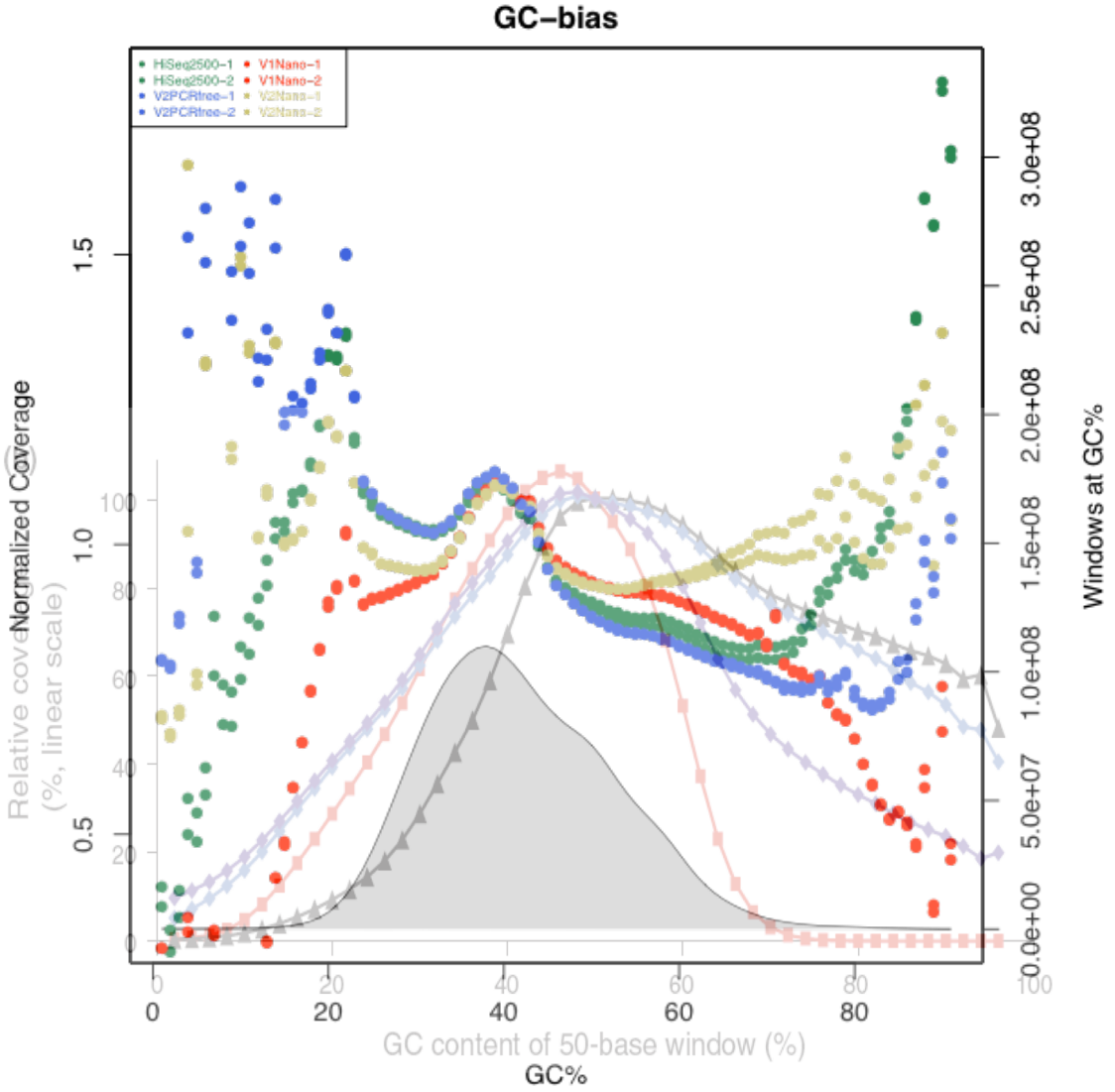
Red = standard PCR protocol
Other colors = modified PCR protocols

From Aird et al., Genome Biology *(2011)*

PCR-Free Libraries

- No PCR amplification in library construction
  - Not the same as no PCR, depending on other steps
- More uniform coverage by GC
- Fewer regions of 0 coverage
- Still some bias as cluster formation is PCR-like
- Requires more DNA (1-2 µg *vs*. 100-200 ng)

GC Bias on Modern Illumina



Sequencing of primers, adapters & tags

- Not every base you sequence is useful
- Primers will be present if you used PCR to target your input DNA
  - Sequence from primers does not represent target
  - Variation seen (or not) under primers is not real
  - Overlapping products will allow analysis of the primer-covered regions
- Short fragments may read through to adapter
- Custom barcodes or other tags may get sequenced too, though most vendor tags will be removed automatically

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing
- General Study Design Considerations
- Considerations for Specific Sequencing Assays

Considerations before starting a sequencing experiment

- What is the question you want to answer?
- How do you decide how much data to generate to answer your question?
  - Sensitivity (e.g. number of False Negatives)
  - Specificity (e.g. number of False Positives)
  - Cost
- Which factors influence the amount of data you generate?

What is the question you want to answer?

- What scientific result do you want?
- Is there an hypothesis you want to test?
  - Early sequencing was “hypothesis free” (i.e. the genome was the goal)
  - Now, it is affordable to sequence for a specific aim (i.e. What sequence do you need for that aim?)
- Understanding this shapes many decisions in designing the experiment

Why choose one type of read?

- Fragments
  - Fastest runs (one read per fragment), least cost
  - Some technologies only make one read
- Paired reads
  - More data per fragment
  - Help with assembly and alignment
  - Same library steps as fragments, but yields more data

Why choose one type of read?

- Mate Pairs (Jumping Libraries)
  - Advantages over paired ends:
    - Paired end separation limited by fragment size
    - Some platforms can't read second strand of fragment
  - Only way to make long links, which are very useful for:
    - Assembly and alignment across repeats and duplication
    - Identification of large structural variants
    - Phasing of small variants
  - Drawback: Requires much more input DNA than paired ends

Why choose one type of read?

- Linked reads
  - Advantages over standard paired ends:
    - Can phase variants over long distances
    - Can be used for assembly scaffolding
    - May aid in single nucleotide and structural variant calling
    - 10X requires low input (1 ng)
  - Drawbacks
    - Additional cost to generate linked read barcoding
    - Requires high quality, high molecular weight input DNA
    - Adds some coverage bias, may require additional coverage
    - May need more coverage to fully utilize haplotype information

Number of reads

- How much data do you need to generate to answer your question?
- This depends on the level of completeness & accuracy you want
- You have to decide before beginning the experiment what level of completeness & accuracy you want, and this determines how much data to generate
- Analogy: Trying a protocol in the lab that requires 1ug of DNA with 0.1ug may end up working, but it may not

Read length

- For most experiments, the longer the reads are the better
- Exception: longer poor-quality reads are not as useful as shorter high-quality reads
- Some experiment types have more stringent requirements for minimum read length

Complexity of library

- Definition of “complexity”: the number of distinct fragments in the library
- After amplification, you may have many copies of the same initial fragment (which does not increase complexity)
- For most experiments, sequencing the same fragment multiple times is not useful and may be detrimental to your analysis

Which sequencing machine to use

- Type of read/library:
  - Illumina & Ion: all
  - PacBio, ONT: fragment
- Read length:
  - Illumina: short (≤150 bp) on HiSeq, medium (≤300 bp) on MiSeq or HiSeq rapid run
  - Ion: medium (200-400 bp, 100-200 for paired end)
  - PacBio, ONT: very long (thousands of bp)

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing
- General Study Design Considerations
- Considerations for Specific Sequencing Assays

Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Example uses of resequencing

- SNP discovery and genotyping
- Population sequencing
- Structural variant discovery and genotyping
- Comparative genomics of closely related species

Considerations before a resequencing experiment

- Considerations for all resequencing experiments
  - Working with a reference genome
  - Aligning reads to a reference
  - Alignability
  - Read length and type
- Considerations for specific types of resequencing experiments
- Targeted resequencing

Working with a reference genome

- How good is the reference?
  - Completeness
  - Accuracy
- How representative is it of your genome(s)?
- Sequence won't align if
  - Absent from the reference
  - Too diverged from the reference

Alignability

- Not all of the reference will be useful for alignment because some parts are too similar for unique alignments (duplications, recent repeats, gene families)
- Longer reads and pairing increase alignability
- Example from human genome resequencing:

	No pairing	400 bp pair	6000 bp pair
36 bp read	85%	96%	-
100 bp read	93%	97%	98%

Adapted from The 1000 Genomes Project Consortium, Nature *(2010)*

Read length and type

- Read length matters for alignability
- Paired end reads also help with alignment
  - Aligning one end uniquely localizes other end
  - Aligners may use this to run more sensitive alignments
  - Allows finding highly variant regions and small indels if the other read from that pair aligns cleanly
- Paired end reads (or very long reads) are necessary for structural variant discovery and genotyping
- Mate pairs (from jumping libraries) are very useful for structural variant analyses but of relatively little use for SNPs and small indels

Considerations for specific types of resequencing experiments

- SNP discovery and genotyping
- Population sequencing
- Comparative genomics of closely related species

Considerations: Sequencing depth for SNP discovery

Type of Experiment	Coverage Required
Haploid SNPs/divergence	≥ 10 x
Diploid SNPs/divergence	≥ 30 x
Aneuploid/somatic mutations	≥ 50 x
Population sequencing	≥ 200 x

Example: Haploid SNP discovery

- You know there is only one base-pair at each locus, so you should make the majority call
- Assuming a uniform 1% error rate, what is the probability that the majority call from your sequencing is actually right?

Depth of coverage at the locus	% of time that majority call is correct	% of time there was no majority call	% of time that majority call is an error
1	99.000	0.00	1.00
2	98.010	1.98	0.01
3	99.970	0.00	0.03
4	99.941	0.06	<0.001
5	99.999	0.00	<<0.001

SNP discovery: Adjusting for random sampling

- Previous graph assumed uniform coverage
- What are the probabilities if the reads are theoretically randomly distributed?

Average depth of coverage across genome	% of time that majority call is correct	% of time there was no majority call	% of time that majority call is an error
1	62.475	37.153	0.372
2	85.646	14.075	0.279
3	94.409	5.432	0.158
4	97.786	2.134	0.081
5	99.110	0.851	0.039
8	99.938	0.059	0.004
10	99.987	0.012	<0.001

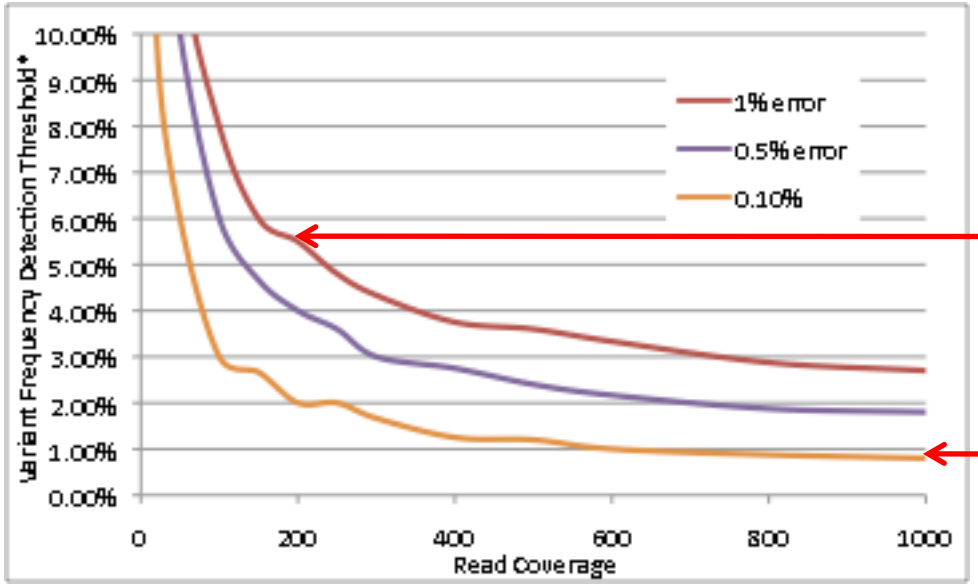
- In reality, distribution will be worse because reads are non-randomly distributed

SNP discovery: Diploid or aneuploid samples

- Diploid samples require twice as much coverage
  - Want to be able to call heterozygotes
  - Need to see each allele as often as you would for a haploid organism
- Aneuploid or somatic mutation samples
  - Cannot rely on expected 1:0 or 1:1 allele ratios
  - Often unique variants, and thus are harder to confirm

Considerations: Population Sequencing

- Example: Want to find all real variants in pooled or host/environmental samples
- What coverage do we need to find a variant at a given frequency?



1% error + 200x
= 5.5% variant

0.1% error + 1000x
= 1% variant

* Lowest frequency of call which exceeds Poisson error probability after Bonferroni correction for 10kb genome

Considerations: Population Sequencing

- Where is the sampling bottleneck?
- Generating more reads than input molecules doesn't improve calling
  - The accuracy and sensitivity of calling is limited by the sampling of the population, not the reads
- With limiting amounts of input, consider using a barcoding scheme that tags input molecules

Considerations: Comparative genomics of closely related species

- Comparative genomic analysis is most effective when species are less than a few % diverged
- Using a more diverged reference:
  - Requires more sensitive (time consuming) algorithms
  - Results in loss of alignability (reads are not placed)
  - Is worse if the divergence is due to insertion/deletion

Targeted sequencing

- Mostly similar to whole genome resequencing
- Targets specific regions (e.g., exome) by:
  - PCR amplification
  - Hybrid selection
  - Targeted genome amplification
- Involves some special analysis considerations

Pros & Cons: Targeted sequencing

- Pros:
  - Significant cost savings if target <<< genome
  - Can achieve higher coverage on target
- Cons:
  - Cost of targeting reagents can be high
  - Some sequenceable regions very hard to target
  - Variability of coverage is higher
  - Targeting may introduce bias
  - Challenging to identify duplicates in targeted sequence

Considerations: Targeted sequencing

- Targeting introduces additional bias
- More coverage required to overcome this (want 3 times or more as much average depth)
- Many off-target reads are generated
  - Not all reads will come from targeted regions
  - Need to bulk up coverage to overcome this
  - Amount will depend on specificity of the targeting

Considerations: Targeted sequencing

- Targeted sequences often include repeats and duplications, and thus some untargeted regions may be sequenced as well
- Need to align to whole genome (not just to the part you targeted) to ensure that unique hits to targeted regions are the best hits for that read in the genome

Considerations: Targeted sequencing

- Some targeting generates identical fragments
  - Examples: PCR targeting, MIPs, HaloPlex
  - Hard to find PCR duplicate reads
  - Many or all starts and ends are the same
- Can use a random barcoding scheme in the amplification to tag fragment of origin
  - Now available with some commercial kits
  - Also refererred to as UMI (unique molecular ids)

Targeted Sequencing Cost/Benefit

- Debate about value of WGS *vs.* Exome
  - Exome provides deeper on target coverage
  - WGS provides more uniform coverage
  - WGS provides whole genome, but can you use it?
  - WGS provides some exonic coverage missed by exome
  - WGS superior for structural variants
- As sequencing costs drop, total costs are coming closer together (3-6x difference)
- Data storage and analysis costs higher for WGS

Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Example uses of genome assembly

- Generate a reference genome
- Alternative method of SNP discovery (even if you have a reference)
  - Mostly for small, haploid genomes
  - Provides better diversity calling for small indels and particularly difficult-to-align regions
- Discover structural variants
  - *De novo* assembly is the only way to get the sequence of a novel insertion
  - Complex structural variants can be more easily discovered through de novo assembly than read alignment to a pre-existing reference

Steps of a genome assembly experiment

- Choose your sample(s)
- Extract DNA from samples
- Fragment the DNA (may need to do this into multiple sizes)
- Library construction (probably need to make multiple libraries)
- Sequencing

Genome assembly considerations: Depth of coverage

- Very deep coverage needed
  - For short reads (Illumina, Ion): 50x – 100x
  - For long reads (PacBio, ONT): 20-50x (required for error correction)
- Common issue is not having sufficient coverage for *de novo* assembly

Genome assembly considerations: Type of reads

- Long reads help greatly
  - Provide connectivity through low coverage
  - Resolve repetitive/duplicated regions
- Paired reads necessary (except w/very long reads)
- Jumping libraries (& mate pair reads) are not always necessary, but yield much better connectivity
- Linked reads also help connectivity and may allow full diploid assembly (e.g., Supernova assembler)
- May use other technologies for very long range scaffolding (Hi-C data, optical mapping, genetic maps)

Genome assembly considerations: Genome complexity and composition

- Repeat content of genome
  - More repetitive genomes require more coverage
  - Paired end reads and jumping libraries more important
- GC content of genome
  - Genomes with extremes of GC content will have more bias in representation
  - Greater average coverage will be required to assemble through extreme GC regions

Genome assembly considerations: Pacific Biosciences

- Highly contiguous assemblies
  - Complete bacteria (with plasmids)
  - Whole eukaryotic chromosome arms
  - Due to long reads and low bias
- Still requires high coverage due to error (~13%)
- Much more expensive than Illumina (~10x)
  - May not matter for small genomes

Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Example uses of RNA-Seq

- Global expression differences
- Annotating genes from a newly sequenced genome
- Discovery of novel genes or transcripts
- Discovery of antisense or other regulatory transcripts
- Variability of isoform expression across conditions

Steps of an RNA-Seq experiment

- Extract RNA from samples
- Enrich for mRNAs
- Make cDNA from RNA
- Fragment the cDNA
- Library construction
- Sequencing

Considerations before an RNA-Seq experiment

- Number of samples needed (conditions and replicates)
- Number of reads needed

Number of samples needed

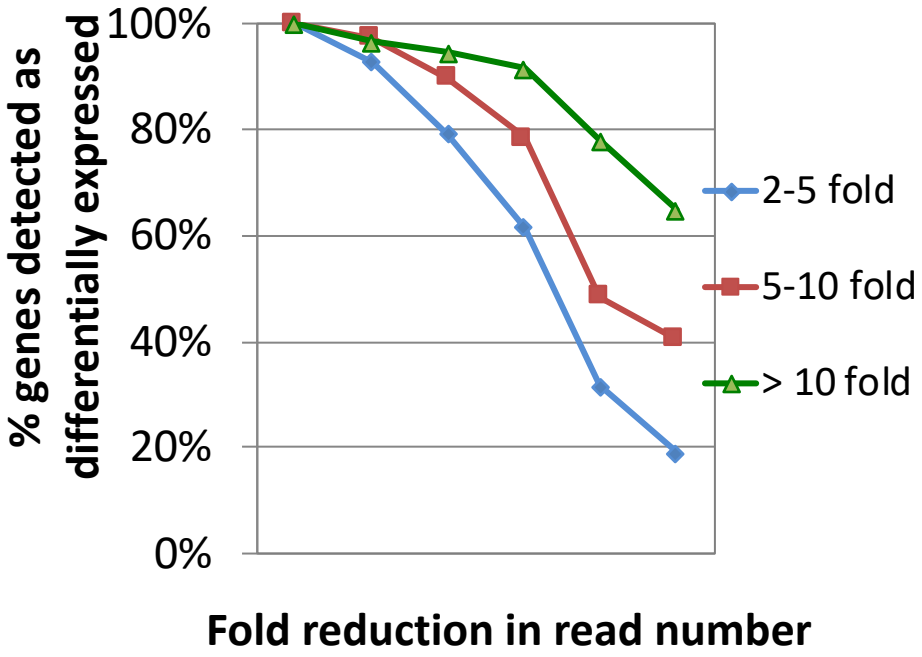
- Number of conditions or tissues determined by experiment:
  - For differential expression, what are you comparing
  - For novel discovery, what are the relevant tissues, conditions, or time points?
- Number of replicates determined by biological variability among replicates
- Website to help estimate optimal power: Scotty
  - http://scotty.genetics.utah.edu/

Number of reads needed

- Need enough reads to identify (and quantify) all transcripts of interest
- How abundant are transcripts of interest?
- What fraction of all transcripts in the cell are in your transcripts of interest?

Number of reads needed

- How large are expression differences?
- Determines significance of the statistical difference



Example RNA-Seq Runs

- Human expression (per condition):
  ¼ lane HiSeq, 75+bp paired (20-60M reads)

  (50-150 transcriptomes per NovaSeq S2, 150-300 S4)
- Vertebrate annotation (per tissue):
  ¼ lane HiSeq, 100+bp paired, strand-specific
- Bacterial and fungal annotation:
  1/12 lane HiSeq, 100+bp paired, strand-specific

Examples of caveats when measuring expression by RNA-Seq

- PCR duplicates don't represent actual counts of RNA fragments, so you need to remove them for quantitation
- Need to be careful about variance:
  - Biological Variance, e.g. Biological variability between replicates of the same conditions may be greater than what is needed to determine statistically significant gene expression changes between conditions
  - Statistical Variance, e.g. When you align reads, they may map to multiple isoforms or multiple paralogs, so you need to assign those reads fractionally to get total transcription levels

Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Example uses of metagenomics

- Characterize species present in an environment
- Determine differences in an environmental population measured at different times or conditions
- Associate metagenomic results with environmental conditions (e.g., host health)

Steps of a metagenomic experiment

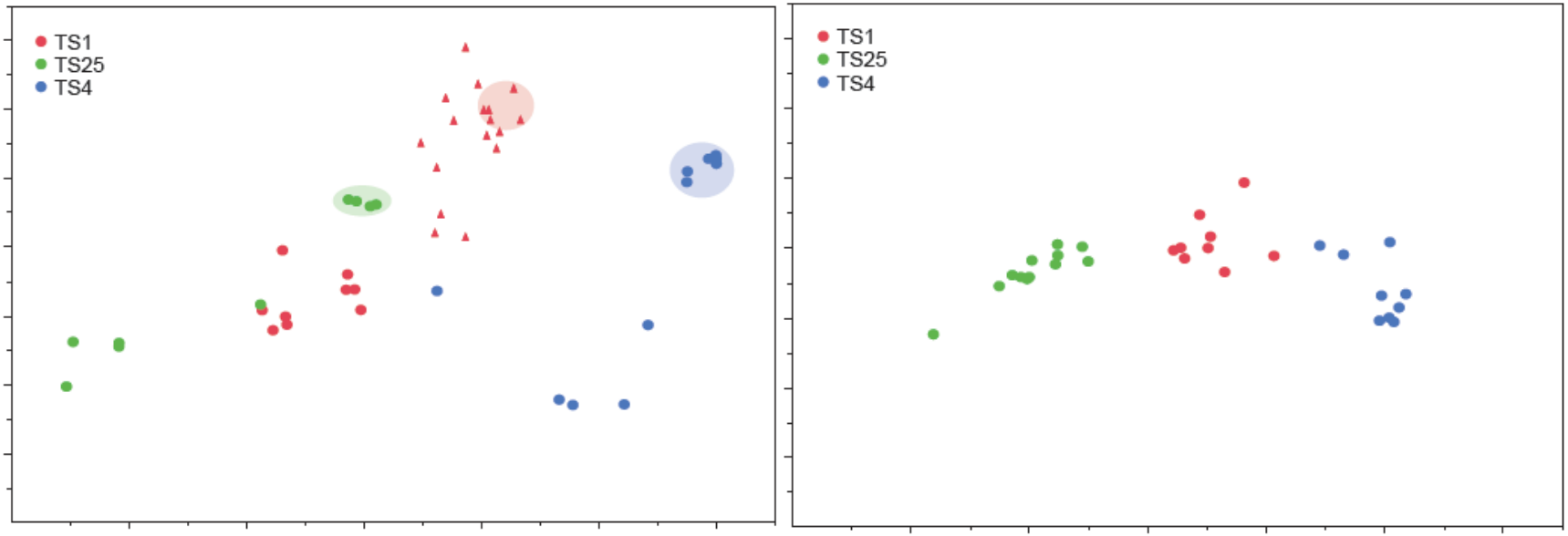
- Extract DNA from samples
- Fragment the DNA (or amplify 16S if not doing whole-genome shotgun sequencing)
- Library construction
- Sequencing

Considerations before a metagenomics experiment

- Reproducibility of metagenomic data depends on:
  - Sample Prep
  - Sequencing Technology
  - Analysis tools
  - Read length and read depth
- Results are not consistent across different experimental designs, but are comparable within identical designs

Metagenomics: Different sample preps

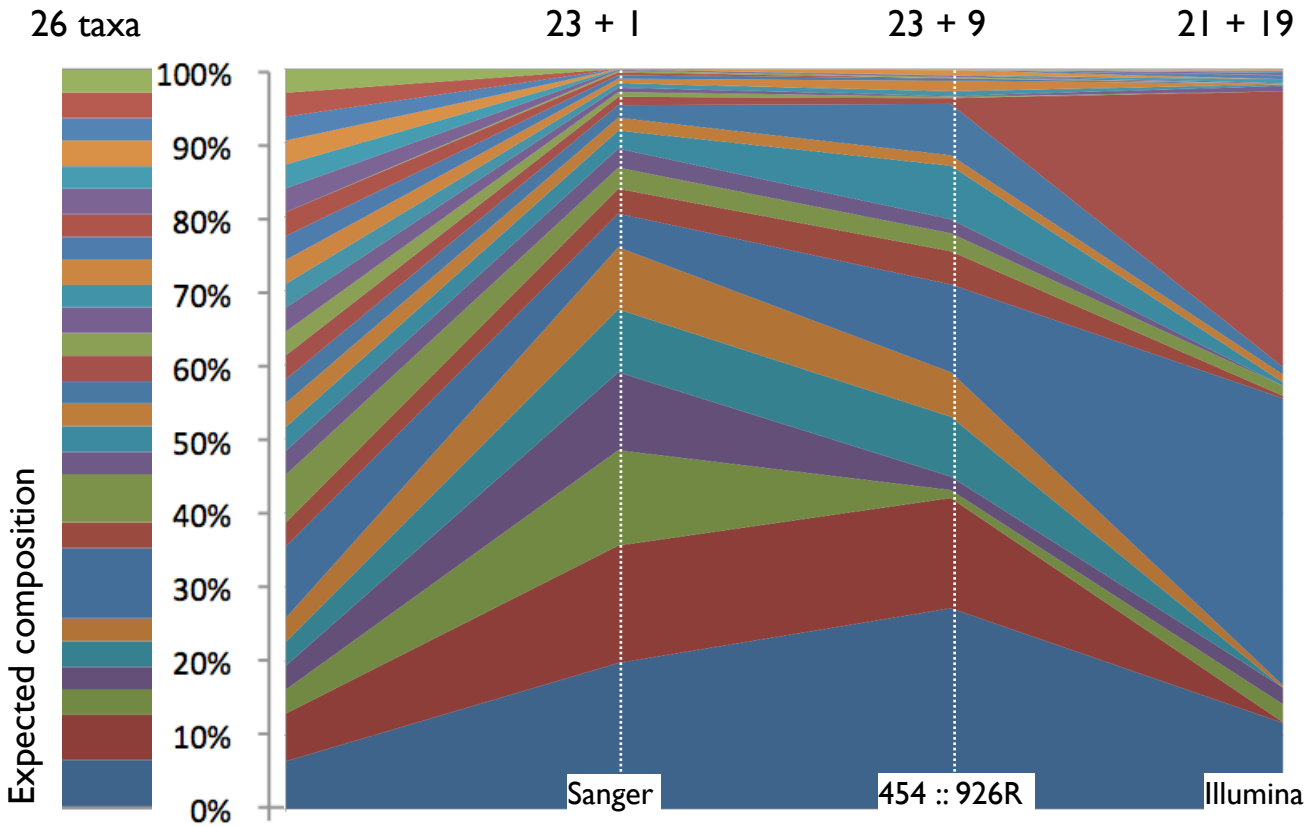
- PCA plots from three samples (colors) sequenced by three groups using different (left) versus identical (right) protocols for sample prep



Human Microbiome Project Data Generation Working Group, submitted

Metagenomics: Different sequencing technologies

- Same (known) mock community sequenced on 3730, 454, and Illumina



From Dirk Gevers

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing
- General Study Design Considerations
- Considerations for Specific Sequencing Assays